WE CLAIM:

1. A method of performing a retrieval operation in a database comprising a tree of nodes, wherein the tree of nodes comprises a root node which is connected to two or more branches originating at the root node, wherein each branch terminates at a node, wherein each node other than the root node may be a non-terminal node or a leaf node, wherein each non-terminal node is connected to two or more branches originating at the non-terminal node and terminating at a node, wherein each leaf node comprises one or more data records of the database, wherein a test associated with each non-terminal node defines a partition of data records based upon one of entropy/adjacency partition assignment and data clustering using multivariate statistical analysis, wherein a current node is initially set to the root node, said method comprising the steps of:

(a) receiving input of a search request providing a retrieval operation and information necessary to perform the retrieval operation;

(b) performing the test associated with a current node responsive to the search request, said test resulting in identification of zero or more distal nodes connected to the current node, wherein said identified distal nodes can, according to the test, contain the data record;

(c) repeating step (b) using an untested distal node which is a non-terminal node as the current node; and

(d) performing the retrieval operation on each identified node that is a leaf node.

2. The method of claim 1, additionally comprising between steps (b) and (c) the step of:

converting an identified leaf node which comprises greater than a threshold number of data records to a new non-terminal node with an associated test, wherein the new non-terminal node is connected to new leaf nodes comprising the individual data records originally stored at the identified leaf node.

3. The method of claim 1, wherein each distal node identified by application of a test to a search request causes the creation of a new search request.

4. The method of claim 1, wherein each branch identified by application of a test to a search request causes the creation of a new search request that is subsequently placed in a search queue from which search requests are subsequently removed for execution by one or more search engines.

5. The method of claim 1, wherein the data records comprise DNA profiles.

6. The method of claim 5, wherein the DNA profiles comprise RFLP data.

7. The method of claim 5, wherein the DNA profiles comprise data on short tandem repeats.

8. A method of partitioning data records in a computer into groups of roughly equal size, comprising the steps of:

(a) defining a function of the probability distribution of the values of a designated variable associated with the data records, wherein the function comprises a linear combination of measures of entropy and adjacency;

(b) partitioning the values of the designated variable into two or more groups, wherein the value of the function is minimized; and

(c) assigning each data record to a group according to the value of the designated variable.

9. The method of claim 8, wherein minimization of the value of the function is achieved by use of a global optimization method.

10. The method of claim 9, wherein the global optimization method produces an approximate result.

11. The method of claim 8, wherein the data comprise DNA profiles.

12. The method of claim 11, wherein the designated variable specifies one or more alleles present at a polymorphic locus.

13. A method of creating a tree-structured index for a database in a computer, wherein the database comprises a tree of nodes; wherein the tree of nodes comprises a root node which is connected to two or more branches originating at the root node, wherein each branch terminates at a node, wherein each node other than the root node may be a non-terminal node or a leaf node, wherein each non-terminal node is connected to two or more branches originating at the non-terminal node and terminating at a node, wherein each leaf node comprises one or more data records of the database, wherein the tree-structured index comprises one or more tests associated with each non-terminal node, said method comprising the steps of:

(a) identifying naturally occurring sets of clusters in the data records of the database;

(b) defining for each identified set of clusters a test that assigns each data record to a cluster within the set of clusters; and

(c) associating each test defined in step (b) with a non-terminal node and an associated set of clusters defined in step (a), and associating with each cluster within the set of clusters one branch originating at the non-terminal node, said branch forming part of one or more paths leading to leaf nodes comprising the data records assigned to the cluster by the test.

14. The method of claim 13, wherein the tests are constructed from entropy/adjacency partition assignments.

15. The method of claim 13, wherein the tests are constructed from clusters identified using multivariate statistical methods.

16. The method of claim 13, wherein the tests are constructed using a combination of entropy/adjacency partition assignment and clusters identified using multivariate statistical methods.

17. The method of claim 16, wherein the tests constructed using clusters identified using multivariate statistical methods are executed by evaluation of a Boolean expression.

18. The method of claim 16, wherein the tests constructed using clusters identified using multivariate statistical methods are executed by evaluation of a decision tree.

19. A method of organizing the data records of a database into clusters, comprising the steps of:

(a) representing one or more variables in each data record in a binary form, whereby the value of each bit is assigned based on the value of a variable;

(b) choosing a set of variables from those represented in all of the data records, whereby principal component analysis of the set of variables yields distinct clusters of the data records;

(c) applying principal component analysis to a sample of the data records, whereby two or more principal component vectors are identified, wherein the scores of the sample data records along these vectors form distinct clusters;

(d) formulating a test based on the identified principal component vectors which assigns each data record to a cluster; and

(e) performing the test formulated in step (d) on each data record, whereby the data records are organized into clusters.

20. The method of claim 19, wherein the value of each bit is assigned in step (a) based on whether the value of a variable is within a designated range of values.

21. The method of claim 19, wherein the value of each bit is assigned in step (a) based on whether a designated value of a variable is present.

22. The method of claim 19, wherein step (c) is performed on a sample of data records from a different database.

23. The method of claim 19, wherein the test formulated in step (d) comprises:

projecting a data record onto the identified principal component vectors;

scaling the projected values;

calculating a distance from the vector of scaled projected values to a representative sample vector of each cluster; and

assigning the data record to the clusters associated with the least distance.

24. The method of claim 23, wherein the representative sample vector of each cluster is the cluster center.

25. The method of claim 19, wherein the data comprise DNA profiles.

26. The method of claim 25, wherein the represented variables are alleles at two or more polymorphic loci.

27. The method of claim 26, wherein the value of each bit is assigned in step (a) based on whether a designated allele is present.

28. The method of claim 19, wherein the represented variables and identified principal component vectors are chosen to yield distinct clusters of approximately equal size.

29. The method of claim 19, wherein each test defines a partition of data of the database according to one of entropy/adjacency partition assignment or data clustering using multivariate statistical analysis.

30. A parallel data processing architecture for search, storage, and retrieval of data responsive to queries, comprising:

a root host processor, responsive to client queries, for creating a search client object and establishing an initial search queue for a query;

a plurality of host processors accessible by said root host processor, each of said root and host processors maintaining a list of available host processors, query queue length, and processing capacity for each processor;

a bus system coupling said host processors; and

a memory for storing a database tree comprising nodes and data of a database accessible via said nodes,

said processors capable of executing a set of tests, associating one test with each non-terminal node of a database tree,

31. A method for search, storage and retrieval of data from a database, comprising the steps of:

defining a set of tests;

associating one test with each non-terminal node of a database tree, each test for defining a partition of data of the database according to one of entropy/adjacency partition assignment or data clustering using multivariable statistical analysis; and

outputting a test result in response to a query by evaluation of one of a Boolean expression or a decision tree.

32. A method of organizing the data records of a database into clusters, comprising the steps of:

(a) representing one or more variables in each data record in a binary form, whereby the value of each bit is assigned based on the value of a variable;

(b) choosing a set of variables from those represented in all of the data records, whereby multivariate statistical analysis of the set of variables yields distinct clusters of the data records;

(c) applying multivariate statistical analysis to a sample of the data records, whereby two or more vectors are identified, wherein the vectors of inner products of the sample data records with the identified vectors form distinct clusters;

(d) formulating a test based on the identified vectors which assigns each data record to a cluster; and

(e) performing the test formulated in step (d) on each data record, whereby the data records are organized into clusters.

33. The method of claim 32, wherein the value of each bit is assigned in step (a) based on whether the value of a variable is within a designated range of values.

34. The method of claim 32, wherein the value of each bit is assigned in step (a) based on whether a designated value of a variable is present.

35. The method of claim 32, wherein step (c) is performed on a sample of data records from a different database.

36. The method of claim 32, wherein the test formulated in step (d) comprises:

projecting a data record onto the identified vectors;

scaling the projected values;

calculating a distance from the vector of scaled projected values to a representative sample vector of each cluster; and

assigning the data record to the clusters associated with the least distance.

37. The method of claim 36, wherein the representative sample vector of each cluster is the cluster center.

38. The method of claim 32, wherein the data comprise DNA profiles.

39. The method of claim 38, wherein the represented variables are alleles at two or more polymorphic loci.

40. The method of claim 39, wherein the value of each bit is assigned in step (a) based on whether a designated allele is present.

41. The method of claim 32, wherein the represented variables and identified vectors are chosen to yield distinct clusters of approximately equal size.